MetaSleepLearner: A Pilot Study on Fast Adaptation of Bio-signals-Based Sleep Stage Classifier to New Individual Subject Using Meta-Learning

Nannapas Banluesombatkul, Pichayoot Ouppaphan, Pitshaporn Leelaarporn, Payongkit Lakhan, Busarakum Chaitusaney, Nattapong Jaimchariyatam, Ekapol Chuangsuwanich, Wei Chen, *Senior Member, IEEE*, Huy Phan, *Member, IEEE* Nat Dilokthanakul* and Theerawit Wilaiprasitporn*, *Member, IEEE*

***Abstract*— Identifying bio-signals based-sleep stages requires time-consuming and tedious labor of skilled clinicians. Deep learning approaches have been introduced in order to challenge the automatic sleep stage classification conundrum. However, the difficulties can be posed in replacing the clinicians with the automatic system due to the differences in many aspects found in individual bio-signals, causing the inconsistency in the performance of the model on every incoming individual. Thus, we aim to explore the feasibility of using a novel approach, capable of assisting the clinicians and lessening the workload. We propose the transfer learning framework, entitled MetaSleepLearner, based on Model Agnostic Meta-Learning (MAML), in order to transfer the acquired sleep staging knowledge from a large dataset to new individual subjects (source code is available at https://github.com/IoBT-VISTEC/MetaSleepLearner). The framework was demonstrated to require the labelling of only a few sleep epochs by the clinicians and allow the remainder to be handled by the system. Layer-wise Relevance Propagation (LRP) was also applied to understand the learning course of our approach. In all acquired datasets, in comparison to the conventional approach, MetaSleepLearner achieved a range of 5.4% to 17.7% improvement with statistical difference in the mean of both approaches. The illustration of the model interpretation after the adaptation to each subject also confirmed that the performance was directed towards reasonable learning. MetaSleepLearner outperformed the conventional approaches as a result from the fine-tuning using the recordings of both healthy subjects and patients. This is the first work that investigated a non-conventional pre-training method, MAML, resulting in a possibility for human-machine collaboration in sleep stage classification and easing the burden of the clinicians in labelling the sleep stages through only several epochs rather than an entire recording.**

***Index Terms*— Sleep stage classification, meta-learning, pre-trained EEG, transfer learning, convolutional neural network.**

This work was supported by PTT Public Company Limited, The SCB Public Company Limited, Thailand Science Research and Innovation (SRI62W1501) and Office of National Higher Education Science Research and Innovation Policy Council (C10F630057).

N. Banluesombatkul, P. Ouppaphan, P. Leelaarporn, P. Lakhan, N. Dilokthanakul and T. Wilaiprasitporn are parts of Bio-inspired Robotics and Neural Engineering Lab, School of Information Science and Technology, Vidyasirimedhi Institute of Science & Technology, Rayong, Thailand (*corresponding authors: natd_pro at vistec.ac.th, theerawit.w at vistec.ac.th).

B. Chaitusaney is with Department of Otolaryngology, Faculty of Medicine, Chulalongkorn University. She is also with Excellence Center for Sleep Disorders, King Chulalongkorn Memorial Hospital, Thai Red Cross Society, Bangkok, Thailand.

N. Jaimchariyatam is with Division of Pulmonary and Critical Care Medicine, Faculty of Medicine, Chulalongkorn University. He is also Head of Excellence Center for Sleep Disorders, King Chulalongkorn Memorial Hospital, Thai Red Cross Society, Bangkok, Thailand.

E. Chuangsuwanich is with the Computer Engineering Department, Chulalongkorn University, Bangkok, Thailand.

W. Chen is with the Center for Intelligent Medical Electronics, Department of Electronic Engineering, School of Information Science and Technology, Fudan University, Shanghai 200433, China

H. Phan is with the School of Electronic Engineering and Computer Science, Queen Mary University of London, United Kingdom.

I. Introduction

A Conventional method of measuring the sleep stages is to conduct the gold standard sleep study called polysomnography (PSG) in sleep laboratory or medical facility [1], [2]. Various sensing techniques, such as electroencephalography (EEG), electro-oculography (EOG), sub-mental electromyography (EMG), electrocardiography (ECG), airflow, etc., are combined to detect the electrical signals emitted from different parts of the human body [3], [4]. During the recording, each 30-second segment of signal interval, i.e., epoch, is annotated by sleep experts to either one of seven stages (Wake (W), Non-Rapid Eye Movement (NREM: S1, S2, S3, and S4), Rapid Eye Movement (REM), and movement time (MT)) following the Rechtschaffen and Kales (R&K) rules [5], or one of five stages (W, REM, and NREM: N1-N3) according to the more recently presented procedure from the American Academy of Sleep Medicine (AASM) [6].

However, the manual scoring, throughout a whole night of sleeping test (approximately 8 hours or 1000 epochs), is time-consuming and burdening the human labor. Therefore, numerous studies have been proposing automatic sleep stage classification by using the main signals containing the characteristics of each stage including EEG, EOG, and sub-mental EMG. The techniques can be categorized into three groups: 1) human-engineered feature extraction with automatic decision making algorithms [7]–[16], 2) the application of extracted features on the Deep Learning (DL) approach [17], [18], and 3) the use of an end-to-end training of the DL approach including Convolutional Neural Network (CNN) or Recurrent Neural Network (RNN) [19]–[25], which involved state-of-the-art sleep staging networks, e.g., DeepSleepNet [26] and SeqSleepNet [27].

Despite the ability to achieve high accuracy of DL model, the method is still not practical due to the massive amount of data needed for training. Furthermore, the models trained on one cohort are not directly applicable to another one due to the data variation from a number of reasons in practice including the amount and the placement of EEG channels, sampling frequencies, experimental protocols, and types of subjects [3], [28], limiting its applicability in clinical environments.

According to aforementioned limitations of the existing DL approaches, one solution for the limited number of samples problem is *transfer learning (TL)* methodology. TL paradigm has a two-step training procedure. The model is pre-trained by a huge dataset, followed by the application (fine-tuning) to the new in-coming dataset [28]. Further elaborations are described in Subsection III-A.

In our previous work, we have shown that it is possible to apply the TL methodology, i.e., pre-training the Auto-encoder, to EEG data in an event-related potential (ERP) classification task [29]. Recently, researchers have also begun to employ TL method for sleep stage classification. Some methods were applied directly from the image classification models [30], while many of them transferred

sleep staging knowledge from large sleep cohorts to other smaller cohorts [31], [32]. However, signals recorded from each subject has a high variation, e.g., sleep pattern, sleep stage duration, and EEG power spectra, etc. [33]. Therefore, the model performs with different accuracy level for different subjects. Thus, fine-tuning is needed to ensure the reliability of the model.

II. Motivation and Contribution

Several works have paved a way to personalize, i.e., adjusting the model to serve for each subject. For example, instead of transferring knowledge from large to small datasets, Mikkelsen *et al.* [34] proposed a personalized model and pre-trained 19 subjects. The pre-training was referred as a generalized model, following by the fine-tuning using the first night of one target subject within the same dataset. The model was evaluated on the individuals data on the second night. Using more diverse data, Andreotti *et al.* [28] transferred knowledge from a combination of two large datasets to other smaller datasets in both healthy subjects and patients. The model was personalized by fine-tuning each subject using 20 patients in the target cohort along with the first night of one target patient. The model was then tested on the second night data of that patient. The limitation of these two works, however, requires the training of the subjects with similar characteristic of sleep stages, i.e., subjects from the same cohort or having similar sleep-related disorders to the target subject. Another study recently published by Phan *et al.* [35], involved the fine-tuning of the pre-trained SeqSleepNet, which was performed on a large dataset, to each subject in another small dataset by using only the first night of each target subject. KL-divergence regularization was applied in order to avoid the over-fitting issue due to the large network though small fine-tuning data. The results showed that personalizing improved the performance on sleep stage classification on each subject's data. However, the data recorded on patients with sleep disorders who have different sleep characteristics are usually more difficult to classify [22], [36] and have not been specifically observed. Furthermore, in many models, the first night of each target subject is still required for the performance.

In a different approach, Chen *et al.* [37] proposed the sleep stage personalizing based on Symbolic Fusion (SF) and Differential Evolution (DE). Firstly, a set of digital parameters based on the domain knowledge of sleep medicine, such as EEG sleep spindle, was extracted from the raw signals. The clinicians labelled sleep stages for only 5% of each record. The DE method was applied to those samples to automatically generate the thresholds, which transformed those digital parameters to symbolic features (e.g., high, medium, and low). The symbolic features were then fed into the inference method in order to classify the sleep stages. Finally, the results were modified with their correction rules based on common patterns of sleep stages. The results were demonstrated to outperform the normal feature extraction methods with ML approach. Still, the classification of N1 stage yielded lower accuracy than the results from DL approach. Aside from the performance, there were also other limitations compared to DL approach: 1) Domain knowledge is required to frame the rules for feature extraction and sleep stage classification which can be done automatically by DL, 2) the performance does not rely on the classification algorithm only, but also the feature extraction method which might not be generalized as the team evaluated on their internal datasets including only 16 subjects, and 3) the decision rules were structured by domain knowledge, therefore, some remainders might not be included due to the variability of the ability to detect installed in the machine. In the learning system, not only the human can teach the machine, but the machine can detect some information insight which might not be definable by humans. Thus, the DL approach appears to be the best method in order to achieve high performance.

In the present work, we presented a pilot study employing a framework, namely *MetaSleepLearner*, to explore the feasibility of solving the aforementioned limitations. Firstly, we used the DL-based approach to develop and possibly expand to achieve the highest performance. We also aimed to solve the issue of limited numbers of data by using TL approach. Secondly, the differences in the recording of each subject were solved by transferring knowledge from pre-trained subjects to new individuals. Thirdly, to mitigate the time-consuming problem from manually labelling the signals recorded the whole night, we motivate and encourage the collaboration between clinicians and machines. The proposed framework allows clinicians to label only several samples per sleep stage, while the remaining labelling can be done automatically by the system. To accomplish our goal, an advanced TL method called Model Agnostic Meta-Learning (MAML) [38], were applied in our approach. To our best knowledge, MAML, which has been widely used in computer visions, has never been used in sleep stage classification. While all previous sleep stage TL works have been focusing on the performance in each state-of-the-art network or different fine-tuning paradigms, we instead attempted to elicit the performance of MAML in the pre-training phase. Its capability includes pre-training on various tasks and is claimed to be an algorithm with fast adaptation to the new tasks by using only a few samples. Therefore, the comparison between the conventional TL method and MAML were mainly investigated. Lastly, the results on both healthy subjects and patients were investigated, in order to ensure the performance on different variations of subjects.

III. Method

A. Transfer Learning in Deep Learning

Many DL researchers and practitioners do not have enough computational resources or sufficient data to train DL models from scratch. It has been found to be more practical to adapt existing models, which are shared in the community, to the task at hand. This practice, namely *Transfer Learning (TL)*, describes heuristics of how to consolidate knowledge from one task (i.e., pre-training) and unpack it in other learning tasks (i.e., adaptation). It has become very influential in many fields, such as robotics [39], [40], computer visions [41], [42], and natural language processing [43]–[45].

B. Model Agnostic Meta-Learning

In this work, we focused on an advanced TL method, called MAML [38]. Normally, TL paradigms compose of 2 phases: pre-training and adaptation (or fine-tuning). MAML describes an algorithm that learns a set of pre-trained weights, which can be easily adapted to new tasks, in contrast to the conventional one, which tries to reach the global optimum from the given data. The benefits of MAML go beyond the reuse of features because, while consolidating, MAML considers the possible changes in these features in the adaptation phase. In other words, MAML grasps (in pre-training phase) at how to quickly learn or adapt in related tasks. We call this ability of learning to learn – *meta-learning*.

This meta-learning ability of MAML is interesting for our task because, unlike image data, the useful features of the bio-signals are less understood. While it is easy to reuse the primitives in image data, it is unclear whether a bio-signal feature from one person (or one device) can be reused, without adaptation, for another person (or another device). Moreover, the MAML itself claims to be a fast adaptation method, which serves our goal, i.e., requiring only a few samples from new individuals for adaptation.

Formally, a neural network, f_θ, is parameterized by its parameters or model's weights, θ. The objective of MAML is to find an initial $\theta = \theta_0$, called meta-weights, that, after updated with stochastic

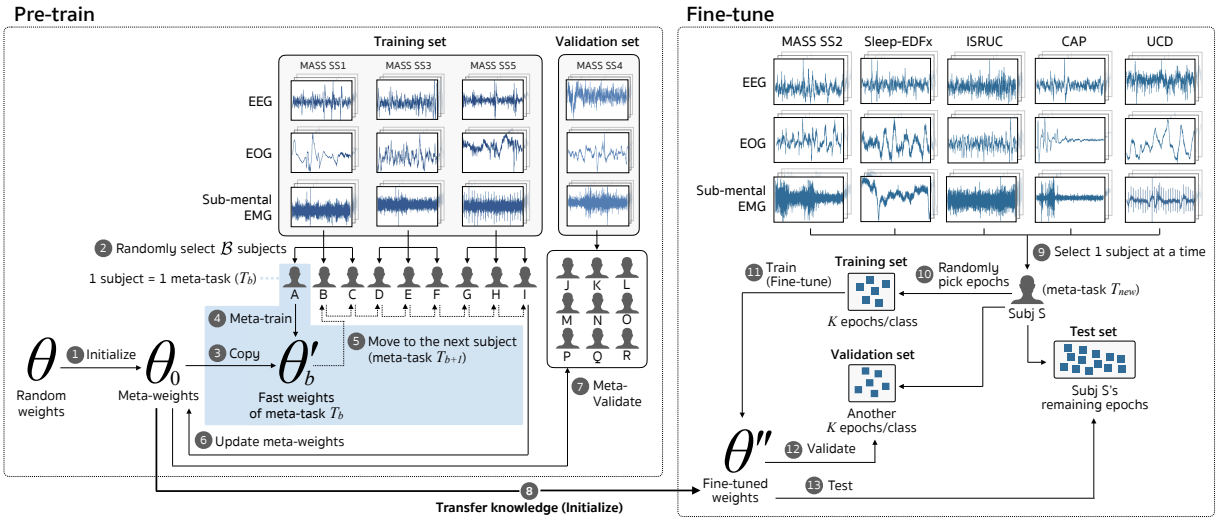

Fig. 1: *MetaSleepLearner*. Our proposed framework composes of two phases: Pre-training & Fine-tuning. The process of pre-training begins by the initialization of the meta-weights, abbreviated as θ_0, from randomized weights to the deep neural network (DNN) model. Then, the meta-training was performed using the data from MASS dataset resulting a pre-trained version of θ_0. The pre-training phase ends when the model acceptably trains the θ_0 to be fitted to the data extracted from the validation set. The second phase of fine-tuning commences to test the performance of θ_0 by transferring the knowledge from the pre-training phase (θ_0) to fine-tune a subject from other cohorts. The model (θ'' initialized by θ_0 from the pre-training phase) is then fine-tuned using the training set and tuned hyperparameters with the validation set. Ultimately, the performance of θ'' is evaluated using the test set, concluding the ability of the fine-tuning phase to be used with any subject with altered conditions.

Algorithm 1: Meta-training for Fast-Adaptation

Require: Meta-training task T_{train}
for each *meta-training iteration* **do**
  Sample training tasks $T_b \in T_{train}, 1 \leq b \leq \mathcal{B}$
  for each *meta-task* T_b **do**
    Copy: $\theta'_b \leftarrow \theta_0$
    Randomly pick K epochs / sleep stage from T_b for 2 sets: $D_p = \{x_i, y_i\}$ and $D_q = \{x_j, y_j\}$
    for $step = 1 \rightarrow num\_updates$ **do**
      Predict using $D_p : \forall i,\ f_{\theta'_b}(x_i)$
      Calculate loss: $\mathcal{L}_{T_b}(f_{\theta'_b})$
      Update fast weights using Equation 2: $\theta'_b \leftarrow \theta'_b - \alpha \nabla_{\theta'_b} \mathcal{L}_{T_b}(f_{\theta'_b})$
      Predict using $D_q : \forall j,\ f_{\theta'_b}(x_j)$
      Calculate loss: $\mathcal{L}_{T_b}(f_{\theta'_b})$
      Keep $\mathcal{L}_{T_b}$ for further calculation: $\mathbb{L}_{T_b}.append(\mathcal{L}_{T_b})$
    end for
  end for
  Sum loss from all meta-tasks using Equation 3: $\mathcal{L}_{\text{meta}}(\theta_0) = \sum_{b=1}^{B} \mathbb{L}_{T_b}(f_{\theta'_b})$
  Update meta-weights using Equation 4: $\theta_0 \leftarrow \theta_0 - \gamma \nabla_{\theta_0} \mathcal{L}_{\text{meta}}(\theta_0)$
end for

gradient descent ($\theta'_b \leftarrow \theta_0$), performs well on a set of tasks $\{T_0, T_1, .., T_B\}$). In other words, it finds θ_0, which returns the sum of the objective values, after updated, as low as possible. Therefore, θ_0 has to be easily adaptable, i.e., it has to be easily updated into a set of good θ'_b.

The main hypothesis assumes that θ_0 will be able to generalize to unseen tasks T_{new}, which are from the same distribution as $T_1, T_2, .., T_B$. Therefore, by training f_θ on a set of related tasks, we consolidated the knowledge of these tasks into θ_0, which could then be easily adapted (or fine-tuned) to an unseen task.

C. MetaSleepLearner

This pilot study aimed to explore the feasibility of employing MAML to perform fast adaptation of sleep staging to new individuals. The overall process, consisting of two phases, is illustrated in Figure 1. The input data included different channels of bipolar EEG, EOG, and submental EMG, which are described in Appendix I. Due to the variability of bio-signals in each recording and each subject, mentioned in section I, the model was expected to learn and adapt to each of them, i.e., fine-tuned to each record of individual subjects. Therefore, one meta-task (referred as T_b) represents one record. Using a large standard sleep dataset ($T_1, T_2, .., T_B$), the model pre-trained with our approach, called *meta-train*, yielded θ_0 as a result. The knowledge, consolidated on the set of weight θ_0, were then transferred to T_{new}, which represented each new individual in other incoming datasets. The details of training steps are described as follows.

1) Pre-train (Meta-training): We randomly selected 3 subsets of MASS [46] for pre-training (SS1, SS3, and SS5) and another one (SS4) for validation. Each subject in each subset was treated as one *meta-task* (T_b), in which the tasks were divided into two groups: for training (T_{train}) and for validation (T_{val}). The model weights, referred to as meta-weights (θ_0), were firstly initialized randomly using Xavier [47], a randomized weight initialization method.

As described in Algorithm 1 and illustrated in Figure 4 (Appendix II), in each meta-training iteration, $\mathcal{B}$ tasks were randomly selected for meta-training ($T_1, T_2, .., T_B \in T_{train}$). In order to adapt to those selected tasks, each T_b copied the weights from θ_0 as its own fast-weights (θ'_b):

$$\theta'_b \leftarrow \theta_0, \tag{1}$$

Subsequently, two sets of samples were randomly selected from T_b, referred to as D_p and D_q, each of which consisted of K epochs per sleep stage, i.e., the variable number of samples per sleep stage. Thus, for five sleep stages, a total of $K \times 5$ epochs were chosen per set. In order to adapt to those tasks, the gradient descent was performed separately for $num\_updates$ steps. Each step starting from using D_p

to update the θ_b' was shown as follows:

$$\theta_b' \leftarrow \theta_b' - \alpha \nabla_{\theta_b'} \mathcal{L}_{T_b}(f_{\theta_b'}) \tag{2}$$

where, α is an updating step size or learning rate ($update\_lr$). The updated θ_b' was evaluated using D_q, referred as $\mathcal{L}_{T_b}$, which would eventually be kept in a list $\mathbb{L}_{T_b}$. After all $\mathcal{B}$ selected meta-tasks were executed, the summation of losses were calculated as:

$$\mathcal{L}_{\text{meta}}(\theta_0) = \sum_{b=1}^{\mathcal{B}} \mathbb{L}_{T_b}(f_{\theta_b'}), \tag{3}$$

which is the objective function of MAML. As a result of a succession of meta-training iteration, θ_0 was updated with the gradient descent of $\mathcal{L}_{\text{meta}}$ using

$$\theta_0 \leftarrow \theta_0 - \gamma \nabla_{\theta_0} \mathcal{L}_{\text{meta}}(\theta_0), \tag{4}$$

where γ is the learning rate of updating meta-weights ($meta\_lr$). Following the re-calculation of θ_0, the meta-validation was performed by sampling the tasks from T_{val}. The sequence procedures abided by the process of meta-training and only for tuning the hyperparameters, e.g. number of meta-training iterations, learning rates, etc. This meta-validation procedure benefits the pre-training process for fast adaptation because the θ_0 is not selected when it immediately performs well with the validation set. Instead, it is selected when it achieves good validation loss after adapting to those data, resulting in the effective adaptation.

2) Fine-tune (Adaptation): This phase is considered the conventional adaptation method in TL. The knowledge from pre-training phase was transferred to this phase by initializing the fine-tuned weights (θ'') with θ_0. The objective of this study is to quickly adapt to new individuals in the new cohorts. Hence, one-night record from one subject was selected at a time from the unseen cohorts, referred as T_{new}. The sleep epochs, i.e. samples of 30-second long of raw signals, extracted from the selected subject, were randomly divided into three different sets: a set of *K* epochs per sleep stage as a training set, another set of *K* epochs per sleep stage for validation, and the remaining epochs for testing. The adaptation procedure was similar to the adaptation in each meta-training iteration. The gradient descent was performed to adapt the θ'' to each individual subject, using only *K* epochs per sleep stage in training set, i.e., only a small amount of sleep stage labels are required. The θ'' was then validated against the other *K* epochs per sleep stage from the validation set in order to find the suitable hyperparameters. Ultimately, the performance of the network was examined using the remaining epochs.

IV. Experiments

To explore the feasibility of employing MAML in this task, we performed the experiments to support our hypothesis as follows. Firstly, we hypothesized that using TL, i.e., the pre-training of the model from large dataset and the fine-tuning on each target subject, should achieve higher performance than training the model with the target subject's data from scratch. Secondly, using multi-modals (EEG, EOG, and sub-mental EMG) in TL would be more advantageous than only EEG signals for sleep stage classification. Thirdly, the main experiment, the efficiency of our pre-training approach should be more efficiently than the conventional approaches in various sleep cohorts of both healthy subjects and patients with sleep disorders. Lastly, the knowledge acquired from our approach should be reasonable and explainable. In all experiments, the procedure was divided into 2 steps: pre-training and fine-tuning. The experimental setups are described in this section.

A. Datasets

Five publicly available datasets were used in our experiments: 1) MASS [46] was published as a common benchmark dataset for sleep research. The performance took place at three different hospital-based sleep laboratories in Canada, recorded from 200 participants, and separated into five subsets. 2) Sleep-EDF [48] from MCH-Westeinde Hospital, Den Haag, Netherlands, consists of two subsets; SC is a study of age effects on sleep in 20 healthy subjects and ST is a study of the effects of the drug temazepam during sleep in 22 patients with mild falling asleep difficulty. 3) The creator of CAP sleep database [49] aimed to investigate the relation between Cyclic Alternating Pattern (CAP) and pathologies in 108 polysomnographic recordings registered at the Sleep Disorders Center of the Ospedale Maggiore of Parma, Italy. 4) ISRUC [50] contains both healthy subjects and subjects with sleep disorders at the Sleep Medicine Centre of the Hospital of Coimbra University (CHUC), which was created for sleep researchers. 5) UCD encompasses 25 subjects with suspected sleep-disordered breathing problem at the Sleep Disorders Clinic, St Vincent's University Hospital, Dublin [51]. Further descriptions of each dataset are provided in Appendix I.

B. Model Specification

The goals of the experiments were not to compare with other network architectures, but mainly to test the hypothesis that our approach (meta-training) would yield better results than the conventional pre-training methods. Therefore, the experiments were executed using an identical network architecture on each training paradigm for a direct assessment. In order to reduce the computational time and resources, we performed the experiments using a simplified version of CNN network, based on the state-of-the-art model, namely DeepSleepNet [26], as described in Appendix III.

C. Pre-training

The experimental setups of the first step of TL, i.e., pre-training, are described in this section. As mentioned earlier in Subsection III-C, for each meta-task T_b, i.e., each subject, K epochs per sleep stage were randomly chosen for training and other K epochs per sleep stage for validation. The subjects, whose sleep stages were less than $K \times 2$ epochs, were filtered out. To satisfy our goal which allows the clinicians to label only a few epochs of a PSG recording, reducing the strain, the K was set as 10. After filtering the number of epochs per sleep stage, there were 30, 37, and 24 subjects from MASS SS1, SS3, and SS5, respectively, as parts of the training set. For particular subsets which segmented each epoch for 20 seconds long, we appended 5-second segments of its preceding and succeeding epochs to make a 30-second epoch. More details of data pre-processing are described in Appendix I. MASS was selected for the pre-training due to its large amount of data compared to the other public datasets. In basic PSG and all acquired cohorts, C3 is one of the most commonly used EEG electrode placements, in accordance with the International 1020 system. Hence, the bipolar C3-A2 EEG electrodes were chosen as samples along with left - right EOG and sub-mental EMG provided by each subset. To compare the performance of our approach against all baselines, the pre-training details of each approach were set as follows:

1) Our proposed Approach (MetaSleepLearner): The meta-training procedures are described in Subsection III-C. The number of meta-tasks selected in each meta-iteration ($\mathcal{B}$) was set to 9. For meta-validation, to be comparable with the number of meta-tasks, 9 subjects with sufficient number of recorded epochs per sleep stage were randomly selected from MASS SS4 and fixed as the representatives in every model run. The model was meta-trained until the increasing

trend of meta-validation loss was observed and the model's weights (meta-weights θ_0) were kept at the best iteration. The set of hyperparameters included the learning rate at the updating meta-weights ($meta\_lr$ (γ) $\in \{10^{-2}, 10^{-3}, 10^{-4}\}$), the learning rate inside the sub-task adaptation ($update\_lr$ (α) $\in \{10^{-2}, 10^{-3}, 10^{-4}\}$), and the number of updating steps ($num\_updates \in$ {5, 10, 15}).

2) Conventional Approach (Baseline-1): The conventional pre-training method was used as a baseline in comparison to our approach. The training procedure involved all samples in the training datasets as they were pooled towards one place. To avoid the class-imbalanced issue, the training data were over-sampled by duplicating under-present sleep stages to achieve an equal number of epochs per sleep stage, in which the validation loss was found to be lower than the non-oversampling one. In each training iteration, the model randomly drew upon the samples to train using mini batches, performing the gradient descent in the same network architecture as our approach. After all mini batches were put through the network, one training iteration ended. The validation was then performed by utilizing all samples from the validation set, which were from the same set of subjects as meta-validation. The model was trained repeatedly until the validation loss did not improve for at least 100 iterations. The best iteration was maintained as θ_0. The sets of hyperparameters included the $learning\_rate \in \{10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}\}$, and $batch\_size \in$ {256, 512, 1024, 2048}.

3) Conventional Approach with one-batch training (Baseline-2): In Baseline-1, since the large amount of pre-training datasets was put into the network, the conventional training approach tended to fit at very early iterations. Moreover, we also altered to select only the same number of training and validating samples in each training iteration in order to compare directly to our approach. The validation samples were also extracted from the same set of meta-validation subjects. However, instead of using all samples from the validation set as in Baseline-1, we randomly selected only $K \times 2$ or 20 epochs per sleep stage from each subject as our approach. The model was also trained repeatedly until the validation loss did not improve for at least 200 iterations and kept the best iteration as θ_0. The sets of hyperparameters were the same as in Baseline-1, with the exception of $batch\_size$, which were from {250, 350, 450}. It was deemed logical to use this set of $batch\_size$ in order to equalize it to our proposed approach by arranging $K \times 9$ per sleep stage = 450. Additionally, the number multiplying to K was also varied ($K \times 5$ = 250 and $K \times 7$ = 350).

In each pre-training paradigm, all combinations of hyperparameters were performed. The set achieving the lowest validation loss was selected subsequently. The model was performed 5 times with the selected hyperparameters, giving rise to 5 sets of θ_0 from MetaSleepLearner, 5 sets of θ_0 from Baseline-1, and 5 sets of θ_0 from Baseline-2.

D. Fine-Tuning

The knowledge from the first phase were then transferred to the fine-tuning phase for adaptation to new individuals in new cohorts. To fine-tune the model, we used four types of weights initialization for comparison: 1) random initialization using Xavier (training from scratch), 2) θ_0 from MetaSleepLearner, 3) θ_0 from Baseline-1, and 4) θ_0 from Baseline-2. The same fine-tuning procedures were applied to all weights initializations in order to compare the quality, i.e., the adaptability, of pre-trained weights from each approach.

Each model was fine-tuned to each individual subject from the unseen datasets composing of both healthy subjects (as reported in their datasets) and patients, including Sleep-EDF, CAP, ISRUC, UCD, and the remaining subset from MASS (SS2). The EEG electrode placements were used differently, as shown in Table I. The main channel was C3-A2. However, due to some records whose C3-A2 channels were not available from the dataset, we used C3-P3 or C4-A1 instead because their characteristics are comparable.

In the fine-tuning phase, the validation set was used to select the most suitable $learning\_rate \in \{10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}\}$, as well as the number of training iterations. The default K and maximum training iterations were set to 5 and 500, respectively. However, in some patient datasets, the performances after evaluating every approach were very poor. Thus, we extended the K to 10 and the maximum number of training iterations to 1000. In order to demonstrate that the model could be applied to an actual arbitrary situation, e.g., the random selection of any epochs labelled by the clinicians, the model randomly selected samples to run for 5 times per indicated subject and per weight initialization. For the comparison between the performances of all pre-training paradigms, the samples, which were randomly picked for training, validating, and testing during the same round from the same subject, were identical.

E. Model Interpretation using Layer-wise Relevance Propagation (LRP)

By applying the proposed fast adaptation procedure, the pre-trained model was adapted to the new incoming subject, i.e., T_{new}, in the fine-tuning phase. The fine-tuning and the validation were performed using only 5-10 epochs per sleep stage. Regardless of the performance, it is necessary to assess whether the method of learning is reasonable. Hence, the application of the Layer-wise Relevance Propagation (LRP) was used for examination. LRP describes a model interpretation method, allowing the reasons for making each decision to be seen, i.e., prediction. It was employed in the previous work to interpret the classification model using the EEG signals [52]. In this study, LRP was employed, revealing the conception of the reasons that model used for each prediction. Initializing from the fine-tuned model (θ''), LRP propagates backward from the output ($f_{\theta''}$) throughout all the layers, reaching the input layer. The $Relevance$ scores (R) was returned, which was found to be devised from the same shape as the input of the original model $(3000, 1, 3)$. Each value of R indicates the reason behind the contribution of each sampling point from the signals to the decision of the model. The R scores was computed as:

$$R_i = \sum_j \frac{a_i w_{ij}}{\sum_{0,i} a_i w_{ij}} R_j, \tag{5}$$

where R_i is a R of neuron i, i and j are two neurons of any consecutive layers, a_i is an activation of neuron i, and w_{ij} is the trained weight (parameter) connecting between neurons i and j.

V. Results

The results elaborated in this section are divided into two parts. Subsection A describes the suitable hyperparameters and the achieved loss values from both the conventional and our proposed pre-training procedures. In Subsection B to E, the averaged evaluation of the performance after fine-tuning (adapting) to new individuals in the new cohorts are reported, as shown in Table I. The number of evaluation results from each cohort, before being averaged, were 5 pre-trained weights $\times$ the number of subjects $\times$ 5 random times, e.g., 450 result samples for SC or 2400 samples for ISRUC (subgroup I). The performance of each experiment, mentioned in this section, was reported as Cohen's Kappa $\pm$ standard errors ($\mathcal{K} \pm$ SE). The General Linear Model with Repeated Measures (α = 0.05, Confidence Interval = 95%) was used to illustrate the significant differences between the results from each approach. It was found that when the number of test

TABLE I: Performance of five sleep stages classification after fine-tuning on each individual in each cohort. The results were averaged from all 5 pre-trained weights per each paradigm × no. of subjects × 5 times of random samples selection. The subjects reported in this table are only those who contained enough epochs for fine-tuning, validating, and testing in their PSG recording.

Dataset	EEG channel	No. of subjects	Overall Accuracy (%)			F1 per class													MF1			Cohen's kappa (κ)		
						W			N1			N2			N3			REM							
			B-1	B-2	ours	B-1	B-2	ours	B-1	B-2	ours	B-1	B-2	ours	B-1	B-2	ours	B-1	B-2	ours	B-1	B-2	ours	B-1	B-2	ours
Healthy																										
1D-CNN																										
EEG only																										
Sleep-EDF (SC)	Fpz-Cz	18	59.0	60.5	**61.4**	53.2	57.7	58.5	16.1	20.2	19.6	64.9	67.3	67.7	65.8	67.8	**70.7**	52.6	54.3	54.8	50.5	53.4	**54.3**	0.446	0.472	**0.482**
2D-CNN																										
EEG only																										
Sleep-EDF (SC)	Fpz-Cz	18	64.9	**73.6**	72.1	59.6	**75.2**	70.0	21.1	28.7	**29.0**	70.5	**79.2**	78.6	74.3	79.2	79.1	60.6	**69.7**	67.5	57.2	**66.4**	64.8	0.536	**0.644**	0.624
EEG, EOG, Submental EMG																										
Sleep-EDF (SC)	Fpz-Cz	18	70.7	70.8	**74.9**	72.2	72.9	**77.1**	30.3	31.6	**35.9**	76.1	75.6	**79.4**	77.7	75.9	**80.2**	64.1	67.1	**71.5**	64.1	64.6	**68.8**	0.606	0.610	**0.662**
ISRUC (Subgroup 3)	C3-A2	10	72.3	70.7	**75.2**	73.1	71.2	**78.9**	39.6	39.2	**43.6**	71.6	69.4	**73.5**	83.2	81.1	83.6	71.1	70.5	**74.9**	67.7	66.3	**70.9**	0.633	0.613	**0.671**
MASS (SS2)	C3-A2 / C4-A1	19	74.7	72.4	**77.3**	68.2	66.7	**73.7**	28.0	28.0	**32.9**	79.9	77.4	**81.6**	83.7	80.7	**84.3**	73.0	71.4	**77.1**	66.6	64.8	**69.9**	0.648	0.618	**0.682**
CAP	C3-A2 / C3-P3 / C4-A1	6	71.3	70.0	**75.1**	69.6	68.2	**75.2**	22.5	22.5	**27.2**	75.8	74.2	**78.7**	81.1	78.5	**82.6**	68.5	68.7	**74.4**	63.5	62.4	**67.6**	0.611	0.596	**0.661**
Patients																										
2D-CNN																										
EEG, EOG, Submental EMG																										
Sleep-EDF (ST)	Fpz-Cz	15	60.7	60.1	**67.1**	54.3	56.6	**61.0**	25.4	27.6	**33.4**	66.6	64.5	**72.8**	74.3	71.7	**78.8**	50.3	52.5	**57.8**	54.2	54.6	**60.8**	0.476	0.471	**0.554**
ISRUC (Subgroup 1)	C3-A2 / C3-M2	96	67.7	65.3	**71.0**	70.7	68.6	**77.2**	40.7	39.6	**44.2**	65.8	62.4	**68.0**	79.1	76.1	**80.2**	65.2	63.6	**69.7**	64.3	62.1	**67.8**	0.577	0.547	**0.618**
CAP (Patients)	C3-A2 / C3-P3 / C4-A1	49	67.9	67.8	**71.4**	68.9	69.4	**74.8**	29.6	30.3	**33.6**	67.1	66.4	**69.8**	79.1	78.6	**80.1**	64.3	65.4	**70.0**	61.8	62.0	**65.7**	0.570	0.570	**0.615**
UCD	C3-A2	22	49.1	50.5	**56.3**	44.1	46.6	**51.1**	24.6	25.9	**27.5**	48.5	49.9	**60.4**	61.7	64.7	**69.0**	34.1	37.8	**42.7**	42.6	45.0	**50.1**	0.345	0.364	**0.429**

The **bold** numbers represent the highest performance among all paradigms with significant difference from others ($p < 0.05$).
B-1 = Baseline-1, B-2 = Baseline-2, ours = our proposed approach

epochs per class was very low, the results were not quite suitable to interpret as one incorrectly predicted sample could highly affect the per-class performance metrics. Therefore, the results in Table I which are mentioned in this section were from the subjects whose samples were sufficiently obtained for fine-tuning, validating, and testing (at least $K \times 3$ epochs per sleep stage). For those subjects whose number of epochs is less than the number mentioned above, the performance is separately reported in Table II in Appendix.

A. Pre-training

In this Subsection, we investigated the results of pre-training phase from all procedures. In Baseline-1, $learning\_rate = 10^{-2}$ with $batch\_size$ = 1024 yielded the best set of hyperparameters, giving the lowest validation loss. From all 5 runs, the validation loss achieved a range of 0.65 to 0.85. Whereas in Baseline-2, $learning\_rate = 10^{-2}$ with $batch\_size$ = 450 were the best set and the achieved validation loss ranged from 0.66 to 0.75. In comparison, using our approach, the best validation loss achieved from $meta\_lr = 10^{-3}$, $update\_lr = 10^{-2}$, and $num\_updates = 10$. However, the results were not as sensitive to these exact values of the learning rates. The best validation loss was 0.46 and the highest was 0.52. The meta-validation procedure, defining whether the weights are suitable for adapting, instigated lower validation loss in comparison to using the conventional methods, when the loss was achieved once the adaptation reached 10 steps.

The average numbers of training iterations, each of 5 runs, from Baseline-1, Baseline-2, and our proposed approach were 8, 1320, and 10162. The averaged computational times were approximately 22 minutes, 6 hours 19 minutes, and 1 day 19 hours 24 minutes, respectively. It is imperative to note that Baseline-1 yielded the lowest training iteration. However, within each iteration, each mini batch contained several sub-iterations until the whole samples were expended. It was deemed usual as our approach required larger amount of iterations due to the fluctuating loss from the meta-training procedure. Since the models weights were updated separately from the 9 meta-tasks before updating the meta-weights in every iteration, the direction of the gradients was more variable. However, both loss values and number of pre-training iterations did not affect the practical usage as the approach could utilize the θ_0 from this phase to adapt to the data from the new incoming cohorts in the fine-tuning phase.

B. Advantages of Transfer Learning

Prior to inspecting the improvement of the proposed pre-training procedures, we examined whether the network trained using TL performed better than the neural network training from scratch. In this experiment, only EEG signals were used. CNN was then changed to 1D CNN, and the input shape was (3000, 1). The recordings from Sleep-EDF dataset, consisting of healthy subjects, were used as a representative to fine-tune in this experiment as it has been commonly used in EEG-only sleep staging studies. We initialized the model from the four paradigms: our proposed approach, Baseline-1, Baseline-2, and the model without the pre-training (randomly initialized using Xavier), in which each of them was performed in a total of 5 runs. As shown in Table I (EEG only (1D-CNN)), while the fine-tuning using Baseline-1, Baseline-2, and our approach yielded MF1 of 50.5, 53.4, and 54.3, respectively, the model without the pre-training phase (not displayed in the Table I) yielded only 18 ± 0.002. In addition, the F1-scores after training from scratch were only 17.84, 8.76, 25.27, 20.82, and 19.47 from W, N1, N2, N3, and REM, respectively. The model, not surprisingly, resulted in poor performance in line with the hypothesis that the deep neural network (DNN) would require to be trained with a large amount of data in order to achieve a high performance. Therefore, the TL paradigm could become one of the solutions for the issue of small available data.

C. Effects of input information on MetaSleepLearner performance

We hypothesized that the information of inputs might affect both performances of TL and the elicitation of our approach. The performances in the fine-tuning phase were compared among the three

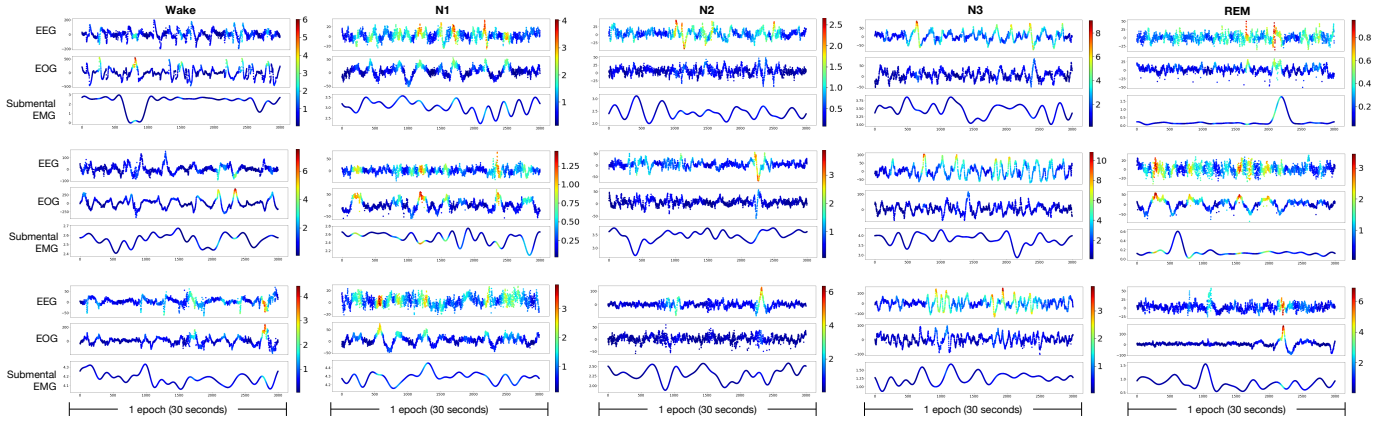

Fig. 2: Visualization results from Layer-wise Relevance Propagation (LRP): Results were sampled from 2 healthy subjects in Sleep-EDF (SC). The prediction of the model on the sleep stages W, N1, N2, N3, and REM are displayed.

types of model structure and inputs: 1) 1D-CNN with EEG, 2) 2D-CNN with EEG, and 3) 2D-CNN with EEG, EOG, and sub-mental EMG. The data from Sleep-EDF (SC) were used in this experiment. As shown in Table I, while using EEG only, after fine-tuning the 2D-CNN from each pre-training paradigm, all of them achieved much better performance than 1D-CNN. When the model structures were 2D-CNN, Baseline-2 surprisingly became worse while being fed with more input data. We also observed this inconsistency when experimenting on a few subjects from other databases (MASS SS2 and UCD) (not shown in the table). That is, some subjects achieved better performance with 3 modalities, while some did not. Moreover, the results from MAML were significantly better than Baseline-1. Although Cohen's kappa obtained by our approach was approximately 0.02 lower than that of Baseline-2 in case of 2D-CNN with EEG only, our approach maintained the best one in case of 3-modality input. This suggests the necessity of using all three modalities in the automatic sleep staging task in this TL setup.

D. Adaptability of MetaSleepLearner in different kinds of subjects

Considering the advantages of MAML, our approach was expected to achieve better performance than the baselines, relying on the conventional TL approaches while enjoyed fast adaptation. In order to provide empirical evidence, we firstly inspected the cohorts of healthy subjects. The number of samples per class used for fine-tuning (K) was set to 5 and the maximum number of training iterations was 500. The first dataset to be explored is the ISRUC (subgroup 3) which used the same EEG electrode placements as the pre-training phase, i.e., C3-A2. The Cohen's Kappa using our approach reached 0.671 $\pm$ 0.005 while using Baseline-1 and Baseline-2 yielded only 0.633 $\pm$ 0.004 and 0.613 $\pm$ 0.005, respectively. The percentage of improvement using our approach was 6.03%, calculated by (MAML the best result of two baselines) / the best result of two baselines, revealing the significant difference ($p < 0.05$) in the mean of all approaches. The second dataset to be explored was CAP, despite containing the data with different EEG channels from the MASS dataset used in the pre-training. Our approach was found to statistically outperform the conventional approaches with 8.28% of improvement in Cohen's Kappa, yielding 0.661 $\pm$ 0.006 compared to 0.611 $\pm$ 0.006 and 0.596 $\pm$ 0.006 of Baseline-1 and Baseline-2, respectively. Similarly, our approach led to an improvement of 5.4% of Kappa on MASS SS2 over the best baseline (Baseline-1), implying the benefit of fine-tuning on the subjects whose recordings were similar to the data used in the pre-training phase. Our approach yielded 0.682 $\pm$ 0.004 while the conventional approaches produced 0.648 $\pm$ 0.004 (Baseline-1) and 0.618 $\pm$ 0.004 (Baseline-2). In a deep inspection of the results, we firstly inspected the cohorts which contained only the subjects who were labelled as healthy. The last dataset, as reported in Subsection V-C, was Sleep-EDF (SC), showing that our approach still performed well and adapted quickly even when a different EEG electrode placements were used. An improvement of 8.48% on accuracy over the best baseline (Baseline-2) was observed. We also showed further results of the cohort from Sleep-EDF (SC) in Appendix, in terms of confusion matrix from all subjects in the cohort and output hypnogram from subject 6. Note that the best baseline, Baseline-2, was used in these comparisons.

To determine whether our approach could be applied to a simulation of a real-world setting, the data of patients with different ratio of each sleep stage and bio-signal characteristics from the control group were used in fine-tuning. Four datasets were chosen to examine the performance of the models adaptation. The results on ISRUC (Subgroup 1) dataset showed that the Cohen's Kappa obtained from all pre-training paradigms was reduced compared to the control groups. Our approach achieved a Cohen's Kappa of 0.618 $\pm$ 0.002 while the Baseline-1 and Baseline-2 attained only 0.577 $\pm$ 0.002 and 0.547 $\pm$ 0.002, respectively. The approach also statistically outperformed the highest performance of conventional methods with 7.05% Kappa improvement. The other cohorts (Sleep-EDF (ST), CAP, and UCD) were also tested. However, we found that the model performed modestly in all TF approaches, i.e., the average MF1 was less than 50. To elucidate the depleted execution, K was increased to 10 epochs per sleep stage. Despite the lower performance (Cohen's Kappa of 0.476 $\pm$ 0.005, 0.471 $\pm$ 0.004, and 0.554 $\pm$ 0.004 by Baseline-1, Baseline-2, and our approach, respectively) on Sleep-EDF (ST) compared to the healthy Sleep-EDF (SC) cohort, our paradigm performed significantly progressing with the improvement of 16.2%. Furthermore, we found that other cohorts required more training iterations. Thus, we extended the maximum of training iterations to 1000 iterations. The results on CAP dataset confirmed the increase with a Cohen's Kappa of 0.615 $\pm$ 0.004. Our approach improved 7.9% over the conventional approach (Cohen's Kappa of 0.57 $\pm$ 0.003). Although the lowest performance was seen on UCD dataset, our approach (Cohen's Kappa of 0.429 $\pm$ 0.006) resulted in an improvement of 17.7% on Kappa over Baseline-2 (Cohen's Kappa of 0.364 $\pm$ 0.006).

Inclusively, using the three modalities from the recordings of both healthy subjects and patients, our approach outperformed the conventional pre-training method with significant differences ($p <$

0.05) in overall accuracy, Cohen's Kappa, and MF1 in almost every sleep stage. This implied that our proposed pre-training paradigm could enhance the performance of TL while enjoying fast adaptation to the new individuals in new cohorts.

E. Model Interpretation

In order to understand what the model learns from our approach, the evaluation by LRP was inspected, as displayed in Figure 2. Three samples per sleep stage recorded from two subjects in the Sleep-EDF (SC) dataset were selected for illustration. The size of the LRP results were the same as the original inputs shape (30 seconds length $\times$ 100 sampling frequency), while the colors represented the level of effects to each prediction by the model, i.e., R scores from LRP. R scores were scaled with all three modalities in each sample. The blue to red colors signified the lowest to highest contribution of the prediction. Only the correct prediction samples were explored in order to determine whether the performance of the model was sufficient with correct learning method.

Figure 2 displays the predictions for W, N1, N2, N3, and REM stages. According to the red-colored sampling points at the left most column figures, EOG signals were found to have an impact on the prediction of W stage. This confirmed the practical views in which the EOG signals would generally show higher activity in the W stage compared to the other stages during the sleep interval. In N1 and REM stages, the similar portions of the three bio-signals were highlighted. This implied the necessity of the three modalities to assess the correct classification. In clinical settings, clinicians routinely identify these stages using all three modalities. EOG signals can be observed to be discriminative between N1 and REM as higher activity is seen in REM [53]. Similarly, the information obtained the reduced activity in sub-mental EMG signals can distinguish N1 from W stage. Furthermore, for stage N2, the model emphasized on the EEG signals, especially the area denoted with the characteristics of K-complex, which are the main EEG traits of this stage. The model also paid attention to only the EEG signals in identifying N3 stage, in which all samples in each epoch were analyzed. In accordance with its name "Slow Wave Sleep", referring to the EEG signals with low frequency, the model required the information of the signal frequency to predict this sleep stage. To that end, the model would require the entire sample length, as illustrated with the non-dark-blue color in the visualization of the EEG modality.

The visualization results revealed that the event occurring inside each epoch displayed the most discriminative information that the model used to classify sleep stages. Moreover, the most discriminative input was found to be the samples in N2 stage, i.e., the samples could be straightforwardly classified as N2 when the K-complex was found. It also verifies that using only several epochs to fine-tune the model is possible for the model adaptation to new individuals in new cohorts as the networks are capable of learning with reasonable predictions and acceptable directions.

VI. Discussion and Limitations

The results confirmed the promise of our approach, outperforming the simplified version of the conventional pre-training methods in both healthy subjects and patients when the bio-signals are present with EEG, EOG, and sub-mental EMG, which are recorded generally in the medical PSG. One explanation for the achievement of the model could be attributed to the meta-learning, which minimizes the losses across all meta-tasks (T_b) after the adaptation of each task. The meta-weights, or θ_0, were generalized while also prompting them to become more adaptable to new tasks or new individuals from newly introduced cohorts with only several samples. In contrast to

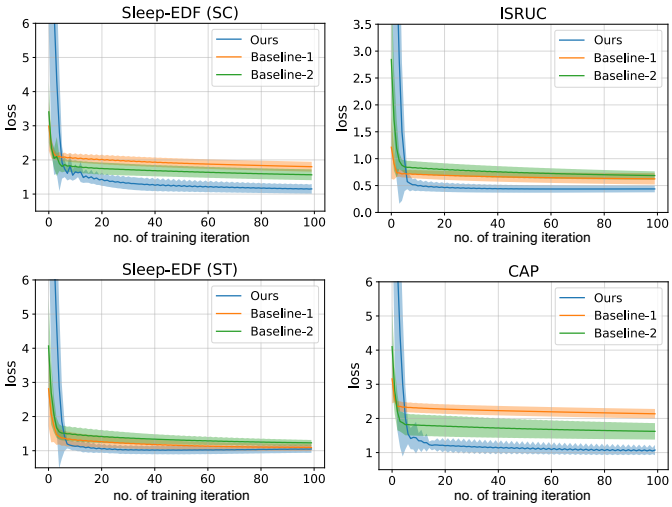

Fig. 3: Fine-tuning loss. Examples of fine-tuning results in which the pre-trained weights from each approach were fine-tuned individually to two healthy subjects (upper) and two patients (lower) from four different cohorts. The plots show average loss with standard deviations of 5 times running while each of them initializes with different pre-trained weights.

the conventional pre-training approach, the model tried to minimize the loss from all samples, making the model more fitted to the training data. To support these statements, some examples of the validation loss during the fine-tuning phase are illustrated in Figure 3. Two healthy subjects from Sleep-EDF (SC) and ISRUC as well as two patients from Sleep-EDF (ST) and CAP were selected for illustration. Only the first 100 iterations were shown, although the model was trained until fitted to the data. The middle line is the mean of the 5 runs in which each run was initialized by different pre-trained weights and fine-tuned on the selected subjects. The areas with colors represent the standard deviations (SD). It can be seen mostly in the results that during the first iteration where fine-tuning was not performed, the θ_0 obtained from our approach acquired a very high loss, which was much higher than using θ_0 from both Baseline-1 and Baseline-2. However, after fine-tuning was performed for only a few iterations, our proposed procedure could fast adapt to the provided data, resulting in much better and eventually higher performances were achieved, as reported in Table I. Granting that the results show that there is feasibility to achieve higher performance using this approach, some limitations still remain in this pilot study, as described in the following section.

A. Adjustment of schematic frameworks

Since our model was modified into a simpler version, to mainly explore the possibility of the proposed approach, it is incomparable to the larger state-of-the-art classification methods. Further adjustments to the base model could be explored in order to improve the effectiveness of the proposed method.

Firstly, the precision rate of sleep stage classification could be increased. The prediction of the stages N1 and REM were shown to be lower than the other stages. One reason might be due to the lower amount of N1 epochs compared to the others. However, this is to be expected as they were in accordance with the general EEG characteristics of both stages. Moreover, in the dataset consisting of the recordings from the patients, the accuracy of every stage tended to decrease. When examining the hypnogram, it is interesting to note that the surrounding epochs might affect the decision and assist in improving the performance. It could be speculated that some of the predictions could be improved when the information regarding

the preceding or the succeeding epochs are presented to the model. Similarly, higher accuracy might be stemmed from an input longer than 30 seconds. Hence, the one-to-many or many-to-one training frameworks from the previous works might affect the balancing of the transitioning sleep stages [21], [24].

Secondly, extending the ability of the networks would unquestionably enhance the classification. In this study, we extracted a modified version of simple CNN networks, imitating from DeepSleepNet [26]. However, it is able to perform with any gradient-based networks. Therefore, the state-of-the-art networks such as DeepSleepNet, SeqSleepNet, or other larger networks could be applied to our proposed TL procedure in order to let the network learn contextual information between epochs using RNNs. To that extent, further exploration using another type of meta-learning which is suitable for RNNs, such as Meta-Learning with memory-augmented neural networks [54], would be accommodating for this task.

Nevertheless, the limitation of MAML involves the requirement of a large amount of computational resources due to the second-order derivation. Thus, the newer method editions, such as iMAML [55] and Reptile [56], might help the evaluation of the larger networks. Although the performance of those methods might not be as good as the original MAML, they require lower amount of resource to operate.

B. Hyperparameter K assessment

K represents the number of training samples per sleep stage. In meta-training, we appointed K = 5, 10, 15 to explore the differences in which more description is available in Supplementary Materials. Not much differences between the results from different K were observed. However, K = 10 produced the highest performance, hence, its application in this study. In the fine-tuning phase, K was set as 5 and 10 alternatively. Although these numbers were chosen in order to regulate the labelling of the samples in a new subject, it should be considered in future works as the exploring of the trade-off between the personality handling and the minimizing number of required labels from clinicians.

C. Standard Hyperparameters assessment

One of important parameters which might affect model's performance is the number of meta-tasks selected in each meta-training iteration ($\mathcal{B}$), which was currently fixed as 9 in our study due to the limitation of computation resources. The another one is the *num_updates* variable, which is the number of updating rounds inside each meta-task during meta-training. In this study, the performances of the model alternating between using 5, 10, and 15 rounds, measured by the validation loss during the pre-training process, exhibited similarities among each other. However, in some conditions, 15 rounds resulted in a higher validation loss than the other rounds. One could speculate that the higher the number of training rounds, the more general interpretation of the knowledge and the further distance from the best parameters of each meta-task (θ_b) could be achieved. Ultimately, although the result suggests that the most optimal *num_updates* for the most fitted validation loss from the three values we selected is 10 rounds, higher or lower values of *num_updates* could still be further explored.

D. Benefits from actual practicality and future works

In order to drive further impact stemmed from our objective of enhancing the human-machine collaboration, it is advantageous for the model to guide the clinicians of which epochs should be labelled, leading to an effective fine-tuning of the model. One approach that could be implemented to enhance the predictive performance is called active learning [57]. It points out which samples, i.e., epochs, benefit the model, in which low confidence in prediction might be exhibited, and presents them to the clinicians to provide their expertise and label the samples efficiently [58].

An alternative way to use learning feedback for improvement is to perform model interpretation. Any improvements will be valuable if the model presents the results and learns in the most possible factual way. Since it is challenging to comprehend the learning contents of the DNNs despite the application of TL, an effectual feedback system is crucial to review the results. The utilization of Layer-wise Relevance Propagation or LRP ensures that the decisions generated by the models stemmed from the same reason the clinicians might use for annotation. One possibility involves the performance metrics showing that the prediction of N1 could be mistakenly classified as REM and vice versa. The visualization states for the reason that it is not only because of the similarity in the characteristics of EEG signals, but also the ways the model learns in both classes. However, our results suggest that the model did not learn to make a reasonably decision in some patients, leading to a lower quality performance in the patients in comparison to the healthy controls. Therefore, the concern over the interpretation of any models along with the adjustment of the other networks or training procedures should be investigated to ensure the precision of the improvement.

The filtering of the subjects with insufficient samples per sleep stage also affected the learning ability of the model. For illustration, we separated those who held samples less than $K \times 3$ epochs per sleep stage to another table, shown in Table II in Appendix. The samples from each subject were selected in the same paradigm as our main experiments, which were selecting K samples per sleep stage to train, another K to validate, and the rest for testing. Therefore, if any subjects contained only a few samples, i.e., less than $K \times 3$, there would be only a few samples left for testing. Additionally, if the number of samples were less than $K \times 2$, the samples would not be enough for validation. However, our proposed approach still mostly outperformed both of the conventional baselines with significant differences ($p < 0.05$).

In addition, the pre-training procedures can be improved by feeding more datasets to facilitate the learning of the pre-trained models from subjects with different demographic and clinical backgrounds. Despite its advantage, data variation from different cohorts may complicate the model processing. Thus, a deep inspection is essential. Moreover, the data from an actual clinical oriented recording (non-public) may have proven important for data diversity. However, more datasets are typically accompanied by longer computational time and more efforts for pre-training. In order to mitigate the problem, the clustering method for grouping similar subjects and the representatives selection instead of using every subject are appealing approaches to be applied in the future.

VII. Conclusion

This pilot study explored a feasibility to perform fast adaptation by applying a Model Agnostic Meta-Learning (MAML) approach, in order to transfer the acquired sleep staging knowledge from a large dataset to new individuals in new cohorts. A simplified edition of the Convolutional Neural Network (CNN) was pre-trained using our approach with MASS dataset, followed by the adaptation or the fine-tuning to each new subject from other cohorts, including Sleep-EDF, CAP, ISRUC, and UCD, by using only several samples from each subject. The performance was compared against the pre-training of two conventional approaches. The investigation using only EEG signals confirmed that only one modality of recordings contained

insufficient knowledge for sleep stage classification. Subsequently, three bio-signal modalities (EEG, EOG, and sub-mental EMG) were employed. One conventional approach displayed poor performance when it retrieved more modalities of input, while both our approach and the second conventional approach achieved higher performance. Moreover, our approach statistically outperformed the conventional approaches as a result from the 5 runs of pre-training and 5 runs of fine-tuning to each individual using the recordings of both healthy subjects and patients. The study also illustrated the learning of the model after the adaptation to each subject, ensuring that the performance was directed towards reasonable learning. This indicates a possibility of using this framework for human-machine collaboration for sleep stage classification, easing the burden of the clinicians in labelling the sleep stages through only several epochs rather than an entire recording.

REFERENCES

[1] J. V. Rundo and R. Downey III, "Polysomnography," in *Handbook of clinical neurology*. Elsevier, 2019, vol. 160, pp. 381–392.

[2] R. K. Malhotra *et al.*, "Polysomnography for obstructive sleep apnea should include arousal-based scoring: an american academy of sleep medicine position statement," *Journal of Clinical Sleep Medicine*, vol. 14, no. 7, pp. 1245–1247, 2018.

[3] R. Boostani *et al.*, "A comparative review on sleep stage classification methods in patients and healthy individuals," *Computer methods and programs in biomedicine*, vol. 140, pp. 77–91, 2017.

[4] O. Faust *et al.*, "A review of automated sleep stage scoring based on physiological signals for the new millennia," *Computer methods and programs in biomedicine*, vol. 176, pp. 81–91, 2019.

[5] D. Moser, P. Anderer, G. Gruber, S. Parapatics, E. Loretz, M. Boeck, G. Kloesch, E. Heller, A. Schmidt, H. Danker-Hopfe *et al.*, "Sleep classification according to aasm and rechtschaffen & kales: effects on sleep scoring parameters," *Sleep*, vol. 32, no. 2, pp. 139–149, 2009.

[6] R. B. Berry, R. Brooks, C. E. Gamaldo, S. M. Harding, C. Marcus, B. V. Vaughn *et al.*, "The aasm manual for the scoring of sleep and associated events," *Rules, Terminology and Technical Sokl,.ppecifications, Darien, Illinois, American Academy of Sleep Medicine*, vol. 176, p. 2012, 2012.

[7] C. Chen *et al.*, "Symbolic fusion: A novel decision support algorithm for sleep staging application," in *Proceedings of the 5th EAI International Conference on Wireless Mobile Communication and Healthcare*. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), 2015, pp. 19–22.

[8] K. A. I. Aboalayon, M. Faezipour, W. S. Almuhammadi, and S. Moslehpour, "Sleep stage classification using eeg signal analysis: a comprehensive survey and new investigation," *Entropy*, vol. 18, no. 9, p. 272, 2016.

[9] X. Li, L. Cui, S. Tao, J. Chen, X. Zhang, and G.-Q. Zhang, "Hyclasss: a hybrid classifier for automatic sleep stage scoring," *IEEE journal of biomedical and health informatics*, vol. 22, no. 2, pp. 375–385, 2017.

[10] A. R. Hassan and M. I. H. Bhuiyan, "Automated identification of sleep states from eeg signals by means of ensemble empirical mode decomposition and random under sampling boosting," *Computer Methods and Programs in Biomedicine*, vol. 140, pp. 201 – 210, 2017.

[11] A. R. Hassan and M. I. H. Bhuiyan, "An automated method for sleep staging from eeg signals using normal inverse gaussian parameters and adaptive boosting," *Neurocomputing*, vol. 219, pp. 76 – 87, 2017.

[12] A. R. Hassan and A. Subasi, "A decision support system for automated identification of sleep stages from single-channel eeg signals," *Knowledge-Based Systems*, vol. 128, pp. 115 – 124, 2017.

[13] A. A. Gharbali, S. Najdi, and J. M. Fonseca, "Investigating the contribution of distance-based features to automatic sleep stage classification," *Computers in biology and medicine*, vol. 96, pp. 8–23, 2018.

[14] E. Alickovic and A. Subasi, "Ensemble svm method for automatic sleep stage classification," *IEEE Transactions on Instrumentation and Measurement*, vol. 67, no. 6, pp. 1258–1265, 2018.

[15] D. JIANG *et al.*, "Robust sleep stage classification with single-channel eeg signals using multimodal decomposition and hmm-based refinement," *Expert Systems with Applications*, vol. 121, pp. 188 – 203, 2019.

[16] T. Zhang *et al.*, "Sleep staging using plausibility score: A novel feature selection method based on metric learning," *IEEE Journal of Biomedical and Health Informatics*, pp. 1–1, 2020.

[17] O. Tsinalis, P. M. Matthews, and Y. Guo, "Automatic sleep stage scoring using time-frequency analysis and stacked sparse autoencoders," *Annals of biomedical engineering*, vol. 44, no. 5, pp. 1587–1597, 2016.

[18] C. Sun, C. Chen, W. Li, J. Fan, and W. Chen, "A hierarchical neural network for sleep stage classification based on comprehensive feature learning and multi-flow sequence learning," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 5, pp. 1351–1366, 2020.

[19] H. Dong, A. Supratak, W. Pan, C. Wu, P. M. Matthews, and Y. Guo, "Mixed neural network approach for temporal sleep stage classification," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 26, no. 2, pp. 324–333, 2017.

[20] S. Chambon, M. N. Galtier, P. J. Arnal, G. Wainrib, and A. Gramfort, "A deep learning architecture for temporal sleep stage classification using multivariate and multimodal time series," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 26, no. 4, pp. 758–769, 2018.

[21] A. Sors *et al.*, "A convolutional neural network for sleep stage scoring from raw single-channel eeg," *Biomedical Signal Processing and Control*, vol. 42, pp. 107–114, 2018.

[22] K. Henri *et al.*, "Accurate deep learning-based sleep staging in a clinical population with suspected obstructive sleep apnea," *IEEE journal of biomedical and health informatics*, vol. 24, no. 7, pp. 2073–2081, 2019.

[23] M. Perslev, M. Jensen, S. Darkner, P. J. Jennum, and C. Igel, "U-time: A fully convolutional network for time series segmentation applied to sleep staging," in *Advances in Neural Information Processing Systems*, 2019, pp. 4415–4426.

[24] H. Phan, F. Andreotti, N. Cooray, O. Y. Chn, and M. De Vos, "Joint classification and prediction cnn framework for automatic sleep stage classification," *IEEE Transactions on Biomedical Engineering*, vol. 66, no. 5, pp. 1285–1296, May 2019.

[25] Q. Wei *et al.*, "A residual based attention model for eeg based sleep staging," *IEEE Journal of Biomedical and Health Informatics*, 2020.

[26] A. Supratak *et al.*, "Deepsleepnet: a model for automatic sleep stage scoring based on raw single-channel eeg," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 25, no. 11, pp. 1998–2008, 2017.

[27] H. Phan, F. Andreotti, N. Cooray, O. Y. Chén, and M. De Vos, "Seqsleepnet: end-to-end hierarchical recurrent neural network for sequence-to-sequence automatic sleep staging," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 27, no. 3, pp. 400–410, 2019.

[28] F. Andreotti, H. Phan, N. Cooray, C. Lo, M. T. Hu, and M. De Vos, "Multichannel sleep stage classification and transfer learning using convolutional neural networks," in *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2018, pp. 171–174.

[29] A. Ditthapron *et al.*, "Universal joint feature extraction for p300 eeg classification using multi-task autoencoder," *IEEE Access*, vol. 7, pp. 68 415–68 428, 2019.

[30] A. Vilamala, K. H. Madsen, and L. K. Hansen, "Deep convolutional neural networks for interpretable analysis of eeg sleep stage scoring," in *2017 IEEE 27th International Workshop on Machine Learning for Signal Processing (MLSP)*. IEEE, 2017, pp. 1–6.

[31] H. Phan, O. Y. Chén, P. Koch, A. Mertins, and M. De Vos, "Deep transfer learning for single-channel automatic sleep staging with channel mismatch," in *2019 27th European Signal Processing Conference (EUSIPCO)*. IEEE, 2019, pp. 1–5.

[32] H. Phan, O. Y. Chén, P. Koch, Z. Lu, I. McLoughlin, A. Mertins, and M. De Vos, "Towards more accurate automatic sleep staging via deep transfer learning," *arXiv preprint arXiv:1907.13177*, 2019.

[33] J. Buckelmüller, H.-P. Landolt, H. Stassen, and P. Achermann, "Trait-like individual differences in the human sleep electroencephalogram," *Neuroscience*, vol. 138, no. 1, pp. 351–356, 2006.

[34] K. Mikkelsen and M. De Vos, "Personalizing deep learning models for automatic sleep staging," *arXiv preprint arXiv:1801.02645*, 2018.

[35] H. Phan, K. Mikkelsen, O. Y. Chén, P. Koch, A. Mertins, P. Kidmose, and M. De Vos, "Personalized automatic sleep staging with single-night data: a pilot study with kl-divergence regularization," *Physiological Measurement*, vol. 41, no. 6, p. 064004, 2020.

[36] R. G. Norman, I. Pal, C. Stewart, J. A. Walsleben, and D. M. Rapoport, "Interobserver agreement among sleep scorers from different centers in a large dataset." *Sleep*, vol. 23, no. 7, pp. 901–908, 2000.

[37] C. Chen *et al.*, "Towards a hybrid expert system based on sleep events threshold dependencies for automated personalized sleep staging by combining symbolic fusion and differential evolution algorithm," *IEEE Access*, vol. 7, pp. 1775–1792, 2019.

[38] F. Chelsea *et al.*, "Model-agnostic meta-learning for fast adaptation of deep networks," in *Proceedings of the 34th International Conference on Machine Learning*, vol. 70. JMLR. org, 2017, pp. 1126–1135.

[39] A. Kai *et al.*, "Classifying options for deep reinforcement learning," in *IJCAI 2016 Workshop on Deep Reinforcement Learning: Frontiers and Challenges*, 2016.
[40] S. James, A. J. Davison, and E. Johns, "Transferring end-to-end visuomotor control from simulation to real world for a multi-stage task," in *Conference on Robot Learning*, 2017, pp. 334–343.
[41] J. Redmon *et al.*, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788.
[42] R. Hu, P. Dollár, K. He, T. Darrell, and R. Girshick, "Learning to segment every thing," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4233–4241.
[43] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.
[44] T. Young, D. Hazarika, S. Poria, and E. Cambria, "Recent trends in deep learning based natural language processing," *ieee Computational intelligenCe magazine*, vol. 13, no. 3, pp. 55–75, 2018.
[45] Y. Zhilin *et al.*, "Xlnet: Generalized autoregressive pretraining for language understanding," in *Advances in neural information processing systems*, 2019, pp. 5753–5763.
[46] C. O'Reilly *et al.*, "Montreal archive of sleep studies: an open-access resource for instrument benchmarking and exploratory research," *Journal of Sleep Research*, vol. 23, no. 6, pp. 628–635, 2014.
[47] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks." in *AISTATS*, ser. JMLR Proceedings, Y. W. Teh and D. M. Titterington, Eds., vol. 9. JMLR.org, 2010, pp. 249–256.
[48] B. Kemp, A. H. Zwinderman, B. Tuk, H. A. Kamphuisen, and J. J. Oberye, "Analysis of a sleep-dependent neuronal feedback loop: the slow-wave microcontinuity of the eeg," *IEEE Transactions on Biomedical Engineering*, vol. 47, no. 9, pp. 1185–1194, 2000.
[49] M. G. Terzano *et al.*, "Atlas, rules, and recording techniques for the scoring of cyclic alternating pattern (cap) in human sleep." *Sleep medicine*, vol. 2, no. 6, pp. 537–553, 2001.
[50] S. Khalighi, T. Sousa, J. Santos, and U. Nunes, "Isruc-sleep: A comprehensive public dataset for sleep researchers," *Computer Methods and Programs in Biomedicine*, vol. 124, 11 2015.
[51] A. L. Goldberger, L. A. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley, "Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals," *Circulation*, vol. 101, no. 23, pp. e215–e220, 2000.
[52] I. Sturm, S. Lapuschkin, W. Samek, and K.-R. Müller, "Interpretable deep neural networks for single-trial eeg classification," *Journal of neuroscience methods*, vol. 274, pp. 141–145, 2016.
[53] M. Ronzhina, O. Janoušek, J. Kolářová, M. Nováková, P. Honzík, and I. Provazník, "Sleep scoring using artificial neural networks," *Sleep medicine reviews*, vol. 16, no. 3, pp. 251–263, 2012.
[54] A. Santoro, S. Bartunov, M. Botvinick, D. Wierstra, and T. Lillicrap, "Meta-learning with memory-augmented neural networks," ser. Proceedings of Machine Learning Research, vol. 48. New York, New York, USA: PMLR, 20–22 Jun 2016, pp. 1842–1850.
[55] R. Aravind *et al.*, "Meta-learning with implicit gradients," in *Advances in Neural Information Processing Systems*, 2019, pp. 113–124.
[56] A. Nichol, J. Achiam, and J. Schulman, "On first-order meta-learning algorithms," *arXiv preprint arXiv:1803.02999*, 2018.
[57] A. Krogh and J. Vedelsby, "Neural network ensembles, cross validation, and active learning," in *Advances in neural information processing systems*, 1995, pp. 231–238.
[58] L. Sheng-Fu *et al.*, "Development of a human-computer collaborative sleep scoring system for polysomnography recordings," *PloS one*, vol. 14, no. 7, p. e0218948, 2019.
[59] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th international conference on machine learning (ICML-10)*, 2010, pp. 807–814.
[60] M. Abadi *et al.*, "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015, software available from tensorflow.org. [Online]. Available: http://tensorflow.org/
[61] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

APPENDIX I

Five publicly datasets were used to evaluate our method. The number of subjects and EEG electrode placements used in each dataset are explained in the Table I and Table II. The study was approved by Rayong Hospital Research Ethics Committee (RYH REC No.E008/2562), Thailand.

A. Montreal Archive of Sleep Studies (MASS)

MASS is a collection of a total of 200 polysomnograms (PSG) from 97 male and 103 female subjects [46]. The cohort was split into five subsets (SS1-SS5), according to the research protocols used when collecting the data. These recordings were manually classified by sleep experts using AASM guidelines with 30-second epoch length for SS1 and SS3 and R&K rules with 20-second epoch length for SS2, SS4, and SS5. The labels following the classification by R&K comprised eight classes, namely sleep stages W, S1, S2, S3, S4, REM, and MT, with those that are unable to be classified labelled as "UNKNOWN [5], [24]. In this study, the AASM guidelines with five sleep stages were followed instead for consistency, merging sleep stages S3 and S4 into N3 and the segments of MT and "UNKNOWN were removed. Furthermore, the EEG and EOG recordings were pre-processed with a notch filter of 60 Hz and band-pass filters of 0.3 - 35 Hz. All recordings from the original sampling frequency of 256 Hz were downsampled to 100 Hz. All segments were generated into 30-second long, extending both ends by 5 seconds for segments that were originally 20-second long.

B. Sleep-EDF

The Sleep-EDF dataset [48] contains two sets of data from two studies: the effects of age on sleep in healthy controls (SC) and the effects of temazepam, a benzodiazepine, on sleep in subjects with difficulty falling asleep (ST). For SC, we used only the same set of subjects as DeepSleepNet [26], such that the number of subjects from SC are 20. For ST, we used all subjects provided from the dataset, which are 22. These recordings had a sampling rate of 100 Hz and were classified according to R&K rules, which were merged into five sleep stages following the AASM standard to be consistent with MASS. Due to the long periods of W stage at the beginning and at the end on most of the recordings, the methods were also modified following DeepSleepNet, by truncating the awake periods at each end to at most 30 minutes. In addition, most subjects contain two-night recorded signals, but only the data obtained during the first night from each subject were used in our experiment.

C. CAP Sleep Database

The title of the Cyclic Alternating Pattern (CAP) Sleep Database is originated from the recurring EEG activity at intervals during NREM [49], where its irregularity may imply a range of sleep disorders. The database consists of 108 PSG recordings, including 16 recordings from healthy subjects and 92 pathological recordings, which were categorized into nocturnal frontal lobe epilepsy (40), REM behavior disorder (22), periodic leg movements (10), insomnia (9), narcolepsy (5), sleep-disordered breathing (4), and bruxism (2). However, we found a problem in accessing the data from 1 bruxism subject. Thus, the remaining 91 pathological recordings were used in our experiment. The sleep stages were scored by expert neurologists according to the R&K rules [51], but also modified according to the AASM standard, similarly to the other datasets.

D. ISRUC

A total of 118 PSG recordings, named ISRUC-Sleep dataset [50], was introduced in 2015 by a team at the Sleep Medicine Centre of the Hospital of Coimbra University (CHUC). The dataset includes three separated subgroups of PSG signals from 100 subjects with history of sleep disorders recorded on one data acquisition session, 8 subjects recorded on two different sessions and two different dates, and 10 healthy subjects. We selected only subgroup 1 and 3 to be representatives in our experiment. The sleep stage classification followed the five sleep stages according to the AASM, labelled by two human experts in each session.

E. St. Vincent's University Hospital / University College Dublin Sleep Apnea Database (UCD)

Revised in 2011, the St. Vincent's University Hospital / University College Dublin Sleep Apnea Database, abbreviated as UCD [51], holds the overnight PSG of 25 subjects (21 male and 4 female) with diagnoses of possible sleep-related breathing disorders such as OSA, central sleep apnea, and snoring. Sleep stages were scored by sleep experts, following the R&K rules, into 8 stages: W, S1, S2, S3, S4, REM, Artifact, and Indeterminate. Similar to other datasets, S3 and S4 were merged into one stage in order to comply with AASM standard and the segments Artifact and Indeterminate were removed in this study. The recordings were pre-processed with a notch filter of 60 Hz, followed by the downsampling from the original sampling frequency of 128 Hz to 100 Hz for consistency with the other datasets. The W periods at the beginning and at the end of each recording which were longer than 30 minutes were trimmed to approximately 30 minutes on both ends.

APPENDIX II

The meta-training procedure, which is described in Subsection III-C, could be illustrated as follows:

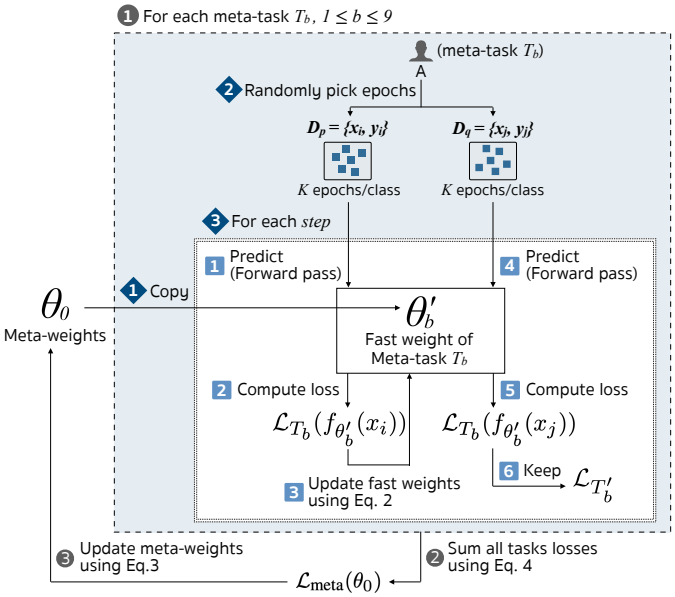

Fig. 4: Diagram of one meta-training iteration.

APPENDIX III

The model, shown in Figure 5, designed as a simpler version of DeepSleepNet, composes of two stacks of CNN layers: small filters (the left/pink stack) and large filters (the right/blue stack), in order to capture the temporal and frequency information, respectively. Each CNN layer is followed by the $relu$ activation function [59]. The $l2$ regularization is used in the first CNN layers as the original network, in order to prevent the over-fitting of noises and artifacts from the signals. However, all CNN layers were changed from 1D to 2D, resembling the study by Phan *et al.* [32] to support multi-modal signals.

For the input data, as described in Appendix I, only MASS dataset was bandpass filtered with the proper frequency range in order to be used as a pre-training data. For other cohorts, all signals were notch-filtered with either 50 or 60 Hz, depending on the datasets and were re-sampled data to reach 100 Hz sampling frequency (if necessary), without any further pre-processing. We assumed that the procedure could handle the raw data from different hardware devices along with different pre-processing methods such as hardware filters.

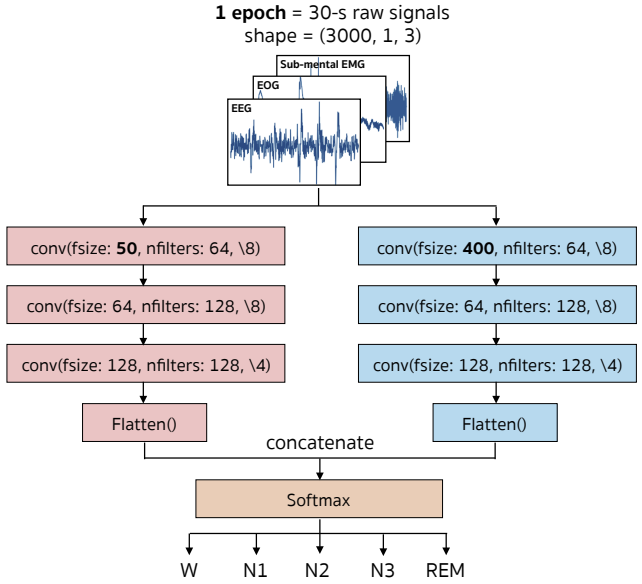

Fig. 5: Example of sleep stage classification network, demonstrating the performance of our proposed fast adaptation procedure. 1 epoch (1 input sample) contains 30 seconds of EEG, EOG, and sub-mental EMG signals. The conv blocks refer to the CNN layers, each with a variety of filter size, number of filters, and stride size, respectively. The bold numbers in the first conv blocks of each side represent the small (left) and large (right) filters of the CNN layers.

The model was implemented using *Tensorflow 1.13.1* [60]. The required inputs in shapes $(width, height, number\_of\_channels)$ could be described as $(no\_of\_samplings, 1, no\_of\_modals)$ = (3000, 1, 3), as the inputs were raw signals from the three modalities. After the inputs were passed through the CNN layers from both large and small filters concurrently, the features extracted from both sides were flatten and concatenated. At the end of the process, one Fully Connected (FC) layer with Softmax activation was used for sleep stage classification. The network was trained using Adam optimizer [61], to minimize the cross entropy loss:

$$\mathcal{L}_{T_b}(f_\theta) = -\frac{1}{N}\left(\sum_{i=1}^{N} y_i \cdot log(f_\theta)\right), \tag{6}$$

where y_i is ground truth of sample i and N is number of samples.

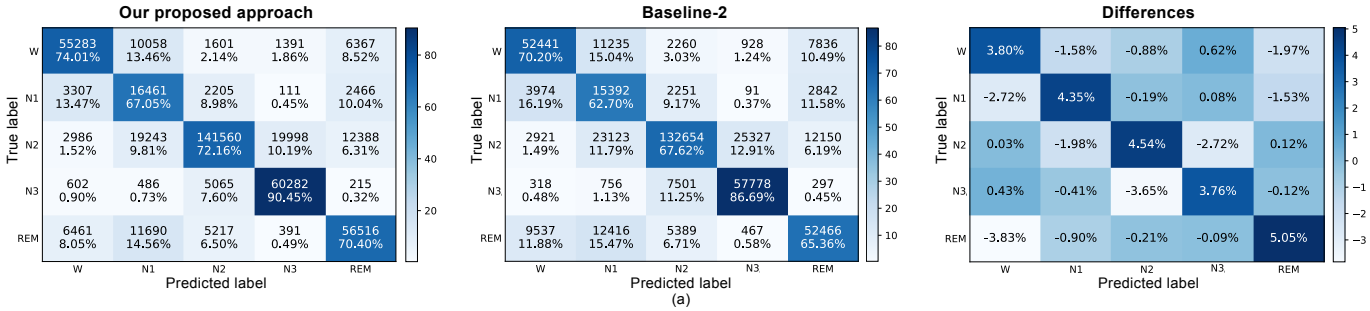

Fig. 6: Examples of confusion matrix after the fine-tuning from pre-trained weights of the proposed approach (left), the Baseline-2 (center), and the differences between both results (right) using (a) Sleep-EDF (SC - Healthy Subjects) and (b) Sleep-EDF (ST - Patients). The results are the summation from all subjects and all runs.

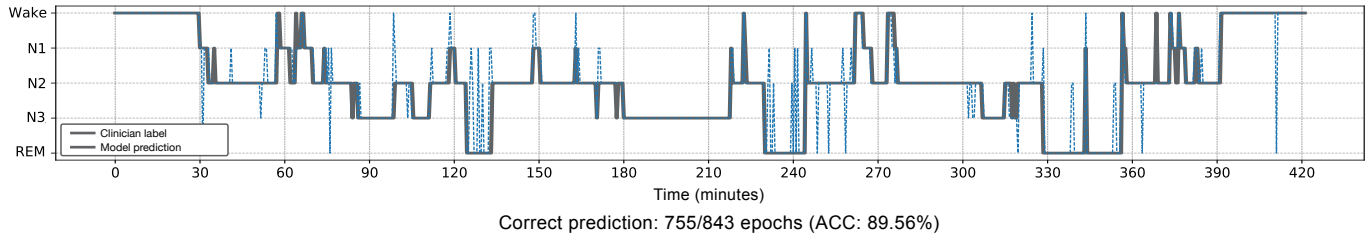

Fig. 7: The hypnogram displaying the comparison between the labelling by clinician and the model's prediction of a representative subject from Sleep-EDF (SC).

TABLE II: Performance of five sleep stages classification after fine-tuning on each individual in each cohort. The results are averaged from all 5 pre-trained weights per paradigm × no. of subjects × 5 times of random samples selection. The subjects reported in this table are only those whose samples are less than $K \times 3$.

Dataset	EEG channel	No. of subjects	Overall Accuracy (%)			F1 per class													MF1			Cohen's kappa ($\mathcal{K}$)		
						W			N1			N2			N3			REM							
			B-1	B-2	ours	B-1	B-2	ours	B-1	B-2	ours	B-1	B-2	ours	B-1	B-2	ours	B-1	B-2	ours	B-1	B-2	ours	B-1	B-2	ours
Healthy																										
2D-CNN EEG, EOG, Submental EMG																										
Sleep-EDF (SC)	Fpz-Cz	2	57.4	54.3	58.6	65.8	61.8	**62.4**	17.8	15.0	18.7	65.2	63.1	67.3	55.5	53.0	55.6	57.1	52.1	56.4	52.2	48.9	51.9	0.407	0.368	0.415
CAP	C3-A2 / C3-P3	7	73.5	69.8	**75.4**	67.5	61.6	**71.8**	11.4	10.4	**13.5**	71.6	68.7	73.1	85.3	81.6	85.5	71.9	68.0	**74.3**	66.8	63.2	**68.8**	0.630	0.580	**0.654**
Patients																										
2D-CNN EEG, EOG, Submental EMG																										
Sleep-EDF (ST)	Fpz-Cz	7	60.1	58.4	**64.9**	35.1	35.5	**38.2**	24.9	23.8	**29.7**	72.8	70.5	**77.1**	52.5	50.5	**54.8**	55.9	57.5	**61.3**	49.2	48.5	**53.4**	0.416	0.401	**0.473**
ISRUC (Subgroup 1)	C3-A2 / C3-M2	4	76.0	75.5	**79.9**	84.5	84.8	**89.4**	26.4	25.9	**31.7**	63.6	59.8	64.9	84.9	84.1	85.8	-	-	-	64.9	63.7	**68.0**	0.629	0.617	**0.685**
UCD	C3-A2	3	48.3	48.8	**53.1**	54.7	58.6	**64.6**	34.1	31.4	32.5	43.1	44.1	**51.4**	85.2	91.1	**93.8**	33.5	34.6	**44.6**	45.7	46.9	**52.4**	0.313	0.329	**0.384**
CAP (Patients)	C3-A2 / C3-P3 / C4-A1	42	68.0	67.2	**70.5**	66.0	65.6	**71.1**	12.0	11.4	**13.4**	65.0	63.4	**66.9**	75.7	74.6	**76.4**	63.7	63.8	**67.3**	63.4	62.7	**66.1**	0.554	0.543	**0.586**

The **bold** numbers represent the highest performance among all paradigms with significant difference ($p < 0.05$).
B-1 = Baseline-1, B-2 = Baseline-2, ours = our proposed approach
"-" (hyphen) designates all 4 subjects from ISRUC (Subgroup 1), which no REM sample remained for testing.